(12) DEMANDE INTERNATIONALE PUBLIÉE EN VERTU DU TRAITÉ DE COOPÉRATION EN MATIÈRE DE BREVETS (PCT)

(19) Organisation Mondiale de la Propriété Intellectuelle
Bureau international

(43) Date de la publication internationale
3 février 2005 (03.02.2005)

PCT

(10) Numéro de publication internationale
WO 2005/010774 A1

(51) Classification internationale des brevets : G06F 17/30

(21) Numéro de la demande internationale :
PCT/FR2004/001930

(22) Date de dépôt international : 21 juillet 2004 (21.07.2004)

(25) Langue de dépôt : français

(26) Langue de publication : français

(30) Données relatives à la priorité :
0308997 23 juillet 2003 (23.07.2003) FR

(71) Déposant *(pour tous les États désignés sauf US)* : FRANCE TELECOM [FR/FR]; 6, place d'Alleray, F-75015 Paris (FR).

(72) Inventeurs; et

(75) Inventeurs/Déposants *(pour US seulement)* : ALLYS, Guillaume [FR/FR]; Arfeuille, F-23460 Royere de Vassiviere (FR). DE BOIS, Luc [FR/FR]; 53, avenue de l'Arche, F-92400 Courbevoie (FR). MARTIN, Stéphane [FR/FR]; 12, rue Littré, F-75006 Paris (FR). KIRSNER, Dominique [FR/FR]; 13, avenue de la Source, F-94130 Nogent-sur-Marne (FR).

(74) Mandataire : BONNET, Michel; Cabinet Lhermet La bigne & Remy, 191, rue Saint-Honoré, F-75001 Paris (FR).

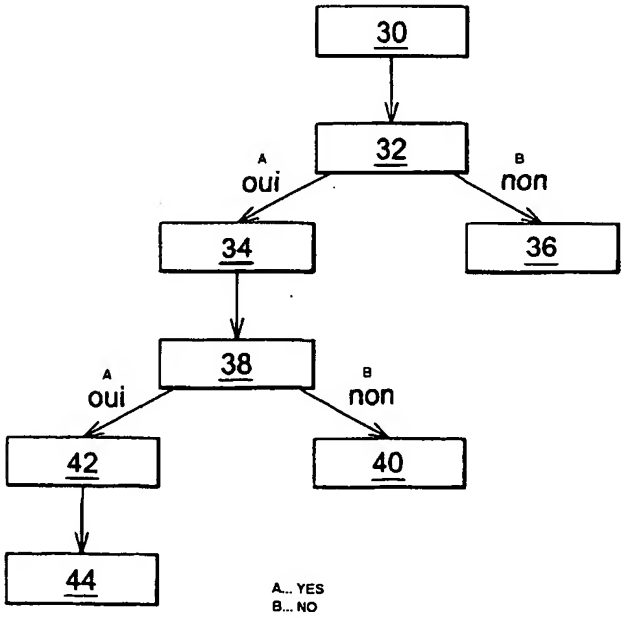
(81) États désignés *(sauf indication contraire, pour tout titre de protection nationale disponible)* : AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) États désignés *(sauf indication contraire, pour tout titre de protection régionale disponible)* : ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,

[Suite sur la page suivante]

(54) Title: METHOD FOR ESTIMATING THE RELEVANCE OF A DOCUMENT WITH RESPECT TO A CONCEPT

(54) Titre : PROCEDE D'ESTIMATION DE LA PERTINENCE D'UN DOCUMENT PAR RAPPORT A UN CONCEPT



(57) Abstract: The inventive method for estimating the relevance of a document with respect to a concept consists in calculating a relevance function thereof with respect to said document on the base of the knowledge of a predetermined semantic neighbourhood of the concept. Said method also involves the calculation (42) of an ambiguity function of said concept in the document, which is distinct from the relevance function, said calculation being estimated in relation with different meanings of the concept in said document. The method is a successor of a preceding step for detecting ambiguous concepts in a knowledge base.

(57) Abrégé : Ce procédé d'estimation de la pertinence d'un document par rapport à un concept comprend le calcul (32) d'une fonction de la pertinence du concept par rapport à ce document s'appuyant sur la connaissance d'un voisinage sémantique prédéterminé de ce concept. Il comporte en outre le calcul (42) d'une fonction d'ambiguïté de ce concept dans ce document, distincte de la fonction de pertinence, ce calcul étant estimé à partir de la présence dans le document de différents sens de ce concept. Ce procédé fait suite à une étape préalable de détection des concepts ambigus dans une base de connaissances.

WO 2005/010774 A1

ZW), eurasien (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), européen (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Publiée :

— *avec rapport de recherche internationale*

— *avant l'expiration du délai prévu pour la modification des revendications, sera republiée si des modifications sont reçues*

En ce qui concerne les codes à deux lettres et autres abréviations, se référer aux "Notes explicatives relatives aux codes et abréviations" figurant au début de chaque numéro ordinaire de la Gazette du PCT.

Procédé d'estimation de la pertinence d'un document par rapport à un concept

La présente invention concerne un procédé d'estimation de la pertinence d'un document par rapport à un concept.

Un procédé classique d'estimation de la pertinence d'un document par rapport à un concept comprend le calcul d'une fonction de pertinence du concept par rapport à ce document s'appuyant sur la connaissance d'un voisinage sémantique prédéterminé de ce concept.

On appelle voisinage sémantique d'un concept, un ensemble de concepts reliés à ce concept par différents liens sémantiques dans une base de connaissances.

En général, lorsqu'on calcule la pertinence d'un document par rapport à un concept, la fonction calculée prend en compte dans son estimation la présence dans le document du concept lui-même, ainsi que celle de tous les concepts appartenant à son voisinage sémantique.

Par conséquent, le résultat d'une requête d'estimation d'un document par rapport à un concept peut être erroné lorsque ce concept est ambigu, c'est à dire lorsqu'il comporte plusieurs sens distincts. En effet, dans ce cas, le voisinage sémantique du concept comporte des concepts voisins de sens différents de ce concept.

Cette ambiguïté est parfois prise en compte dans le calcul de la fonction de pertinence, en réduisant le résultat obtenu par l'estimation de la présence du concept pris dans un sens prédéterminé par un résultat obtenu par l'estimation de la présence de concepts pris dans un sens différent. Ainsi, par exemple, un document dans lequel la présence de concepts pris dans un sens différent est supérieure à la présence de concepts pris dans le sens prédéterminé n'est plus considéré comme étant pertinent par rapport au concept.

Ce type de procédé prenant en compte l'ambiguïté du concept risque donc de considérer un document pouvant intéresser l'utilisateur comme ayant une mauvaise pertinence par rapport à ce concept, par exemple au cas où une fausse détection d'ambiguïté adviendrait.

L'invention a pour but de remédier à cet inconvénient en fournissant un procédé d'estimation de la pertinence d'un document par rapport à un concept capable de prendre en compte l'ambiguïté du concept sans dégrader l'estimation de la pertinence du document par rapport au concept.

A cet effet, l'invention a pour objet un procédé d'estimation de la pertinence d'un document par rapport à un concept, comprenant le calcul d'une fonction de pertinence du concept par rapport à ce document s'appuyant sur la connaissance d'un voisinage

sémantique prédéterminé de ce concept, caractérisé en ce qu'il comporte en outre le calcul d'une fonction d'ambiguïté de ce concept dans ce document, distincte de la fonction de pertinence, ce calcul étant estimé à partir de la présence dans le document de différents sens de ce concept.

Ainsi, la prise en compte de l'ambiguïté est décorrélée du calcul de la fonction de pertinence. La pertinence du document reste donc inchangée en cas d'ambiguïté, et c'est un score déterminant uniquement l'ambiguïté qui avertit l'utilisateur du fait que le document est susceptible de l'intéresser ou non.

Dans le cas d'une fausse détection d'ambiguïté, le document sera toujours considéré comme pertinent par rapport au concept, puisque seul le score déterminant l'ambiguïté est susceptible d'être erroné.

Un procédé selon l'invention peut en outre comporter l'une ou plusieurs des caractéristiques suivantes :

- la fonction de pertinence mesure la présence du concept et des concepts du voisinage sémantique de ce concept dans le document ;
- le voisinage sémantique du concept comporte plusieurs nuages sémantiques de sens distincts, et la fonction d'ambiguïté compare la présence de concepts appartenant à un nuage sémantique correspondant à un sens prédéterminé du concept avec la présence de concepts appartenant à des nuages sémantiques différents ;
- la présence de chacun des concepts appartenant aux différents nuages sémantiques est pondérée par un coefficient prédéterminé ;
- le procédé comporte une étape préalable de détection de concepts ambigus, c'est à dire de concepts comportant plusieurs nuages sémantiques de sens différents dans leur même voisinage sémantique ;
- lors de l'étape de détection préalable, deux concepts sont considérés comme ambigus s'ils sont reliés entre eux par au moins deux liens sémantiques différents.
- lors de l'étape de détection préalable, un concept est considéré comme ambigu s'il est relié à au moins deux nuages sémantiques de sens différents ;
- le concept appartient à une base de connaissance obtenue par fusion d'une première base de connaissances avec une seconde base de connaissances, l'étape préalable de détection des concepts ambigus étant réalisée lors de la fusion.

- lors de l'étape de détection des concepts ambigus, un concept de la première base de connaissances est considéré comme ambigu s'il est relié par un nouveau lien à un autre concept de la première base de connaissances.
- lors de l'étape de détection des concepts ambigus, un concept de la première base de connaissances est considéré comme ambigu s'il est relié à au moins un nuage sémantique de la seconde base de connaissances.

On notera qu'on appelle nuage sémantique d'un concept considéré, un ensemble constitué de concepts reliés à un même sens du concept considéré.

Par exemple, le concept « Orange » comporte dans son voisinage sémantique au moins deux nuages sémantiques de sens différents, à savoir un nuage sémantique se rapportant à la couleur orange (comportant entre autres les concepts « couleur », « jaune », « rouge », etc.) et le nuage sémantique se rapportant au fruit orange (comportant entre autres les concepts « fruit », « agrume », « citron », etc.).

L'invention sera mieux comprise à la lecture de la description qui va suivre, donnée uniquement à titre d'exemple et faite en se référant aux dessins annexés dans lesquels :

- la figure 1 représente schématiquement une base de connaissances constituée de concepts et de liens sémantiques entre eux ;
- les figures 2 et 3 représentent schématiquement une méthode de détection de concepts ambigus, mise en œuvre dans un procédé selon l'invention et ;
- la figure 4 représente schématiquement un procédé d'estimation de la pertinence d'un document par rapport à un concept selon l'invention.

On a représenté schématiquement sur la figure 1 une base de connaissances que l'on désignera par la référence générale 10.

On notera que, dans cet exemple, la base de connaissances 10 est constituée d'une base de connaissances 10A à laquelle on a ajouté une base de connaissances 10B, selon un procédé de fusion de bases de connaissances connu en soit.

Un concept 12 de la base de connaissances 10 est relié à d'autres concepts par des liens sémantiques 14.

L'ensemble des concepts ainsi reliés au concept 12 forme un voisinage sémantique de ce concept 12. Ce voisinage sémantique peut comporter plusieurs nuages sémantiques 16 de sens distincts, un nuage sémantique 16 du voisinage du concept 12 étant, comme cela a été défini précédemment, un ensemble constitué de concepts reliés à un même sens du concept 12 considéré.

Lorsqu'un concept 12 est relié à plusieurs nuages sémantiques 16 de sens distincts, ce concept est dit « ambigu ». Les concepts ambigus sont désignés sur la figure 1 par la

référence générale 18, et par les références particulières 18A, 18B et 18C, ces références particulières correspondant à différents modes de détection des concepts ambigus, mis en oeuvre lors d'une étape préalable d'analyse de la base de connaissances 10. Cette étape sera détaillée en référence aux figures 2 et 3.

Durant cette étape préalable, les concepts possédant plusieurs nuages sémantiques de sens différents dans leur voisinage sémantique sont marqués comme étant ambigus.

La figure 2 représente une mise en œuvre de cette étape préalable, adaptée pour la détection de concepts ambigus dans une base de connaissances donnée, par exemple ici, la base de connaissances 10A.

Chaque concept 12 de la base de connaissances 10A est analysé lors d'une étape 20 durant laquelle on recherche au moins deux liens sémantiques différents reliant ce concept 12 à un seul autre concept.

Dans le cas où ces liens existent, on passe à une étape 21 au cours de laquelle le concept est marqué comme étant un concept ambigu 18A, puisque la présence d'au moins deux liens vers un même autre concept indique une forte probabilité pour que ces liens concernent des sens différents de ce concept.

Dans le cas contraire, on passe à une étape 22 lors de laquelle on recherche au moins deux liens sémantiques reliant ce concept 12 à deux nuages sémantiques de sens différents.

Dans le cas où ces liens existent, le concept est par définition un concept ambigu. On passe alors à une étape 23 lors de laquelle il est marqué comme étant un concept ambigu 18B.

Dans le cas contraire, le concept 12 n'est pas considéré comme étant ambigu, et on passe à une étape 24 de fin d'étape préalable d'analyse de la base de connaissances 10A.

La figure 3 représente une mise en œuvre de l'étape préalable de détection des concepts ambigus, plus particulièrement lors de la fusion de la base de connaissances 10A avec la base de connaissances 10B. Les nouveaux liens créés entre concepts lors de cette fusion sont représentés sur la figure en traits interrompus.

Chaque concept 12 existant dans la base de connaissance 10A est alors analysé lors d'une étape 25 durant laquelle on recherche au moins un nouveau lien sémantique reliant ce concept 12 à un autre concept existant de la base de connaissance 10A, ce nouveau lien ayant été créé lors de la fusion des deux bases 10A et 10B.

Dans le cas où un tel nouveau lien existe, on passe à une étape 26, durant laquelle le concept est marqué comme étant un concept ambigu 18C, puisque la relation entre ces

deux concepts n'était pas prévue dans la base de connaissances initiale 10A, ce qui implique qu'il s'agit potentiellement d'homonymes.

Dans le cas contraire, on passe à une étape 27, durant laquelle chaque concept 12 existant dans la base de connaissance 10A est de nouveau analysé, pour rechercher au moins un lien sémantique reliant ce concept 12 à un nuage de nouveaux concepts de la base de connaissances 10B.

Dans le cas où un tel lien existe, on passe à une étape 28 durant laquelle le concept est marqué comme étant un concept ambigu 18D, puisqu'il est probable que ce lien vers ces nouveaux concepts concerne un homonyme.

Dans le cas contraire, le concept 12 n'est pas considéré comme étant ambigu, et on passe à une étape 29 de fin d'étape préalable d'analyse de la base de connaissances.

Une fois cette étape préalable de recherche de concepts ambigus effectuée, il est possible d'estimer la pertinence d'un document par rapport à un concept donné de la base de connaissances 10, par le procédé représenté schématiquement sur la figure 4.

Lors d'une première étape 30, une requête d'estimation de la pertinence d'un document par rapport à un concept 12 de la base de connaissances 10 est émise, par exemple par un moteur de recherche.

Une fois cette requête émise, on passe à une étape 32, durant laquelle un calcul d'une fonction de pertinence du document par rapport au concept 12 est effectué de manière connue en soi. Cette fonction de pertinence est calculée en prenant en compte la présence dans le document du concept 12 et de concepts appartenant au voisinage sémantique de ce concept 12.

Ainsi, par exemple, la fonction de pertinence est donnée par l'équation suivante :

Pertinence(Doc, 12) = f[Présence(Doc, 12), coef x Présence(Doc, voisinage(12))],

où :

- Pertinence(Doc, 12) est la fonction de pertinence du concept 12 dans le document considéré ;
- Présence(Doc, 12) est une fonction quantifiant la présence du concept 12 dans le document considéré, par exemple, le nombre de fois où le concept 12 apparaît dans le document ;
- Présence(Doc, voisinage(12)) est une fonction quantifiant la présence dans le document considéré de concepts appartenant au voisinage du concept 12 ;
- coef est un coefficient de pondération prédéterminé, permettant de d'accorder plus ou moins d'importance aux concepts appartenant au voisinage sémantique du concept 12 ;

- f est par exemple une fonction « maximum », ou une fonction « somme ».

En fonction du résultat obtenu par ce calcul, le document peut être considéré comme étant pertinent vis à vis du concept 12, par exemple si le calcul donne un résultat supérieur à un seuil prédéterminé. Dans ce cas, on passe à une étape 34 au cours de laquelle le document est marqué comme étant pertinent par rapport au concept 12.

Dans le cas contraire, où le résultat du calcul donne un résultat inférieur au seuil prédéterminé, on passe à une étape 36 au cours de laquelle le document est marqué comme n'étant pas pertinent par rapport au concept 12. Dans ce cas, le document non pertinent n'est pas retenu.

Dans le cas où le document est marqué comme étant pertinent, le procédé selon l'invention prévoit ensuite le calcul d'une fonction d'ambiguïté du concept dans le document.

Lors d'une étape 38, on vérifie si le concept 12 sur lequel porte la requête est marqué comme étant ambigu ou non dans la base de connaissances 10.

S'il n'est pas marqué comme étant ambigu, on passe à une étape 40 au cours de laquelle le document est marqué comme étant pertinent et non ambigu.

Si le concept 12 est marqué comme étant ambigu, on passe à une étape 42 durant laquelle on procède à un calcul de la fonction d'ambiguïté, comparant la présence de concepts appartenant à un nuage sémantique correspondant à un sens prédéterminé du concept 12 (le sens du concept dans la requête) avec la présence de concepts appartenant à des nuages sémantiques différents.

Ainsi, la fonction d'ambiguïté peut être donnée par l'équation suivante :

Ambiguïté(Doc, 12) = f[coef1 x Présence(Doc, nuage1), coef2 x Présence(Doc, nuage2)],

où :

- Ambiguïté(Doc, 12) est la fonction d'ambiguïté du concept 12 dans le document considéré ;
- nuage1 et nuage2 sont deux nuages sémantiques différents reliés au concept 12 considéré ;
- Présence(Doc, nuage1) quantifie la présence de concepts appartenant au nuage 1 dans le document considéré ;
- coef1 est un coefficient prédéterminé, permettant d'accorder plus ou moins d'importance aux concepts appartenant au nuage 1 ;
- Présence(Doc, nuage2) quantifie la présence de concepts appartenant au nuage 2 dans le document considéré ;

- coef2 est un coefficient prédéterminé, permettant d'accorder plus ou moins d'importance aux concepts appartenant au nuage 2 ;
- f est une fonction de comparaison.

Une fois ce score d'ambiguïté calculé, on passe à une étape 44 au cours de laquelle le document, est marqué comme pertinent avec un score d'ambiguïté, et il ne tient ainsi qu'à l'utilisateur d'estimer, à l'aide de ce score d'ambiguïté, si le document est susceptible de l'intéresser ou non.

Il apparaît clairement qu'un procédé d'estimation de la pertinence d'un document par rapport à un concept donné, tel que décrit précédemment, fournit de meilleurs résultats que les procédés existants, en pondérant la pertinence par un calcul d'ambiguïté sans affecter l'estimation de la pertinence elle-même.

REVENDICATIONS

1. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12), comprenant le calcul (32) d'une fonction de pertinence du concept (12) par rapport à ce document s'appuyant sur la connaissance d'un voisinage sémantique prédéterminé de ce concept (12), caractérisé en ce que, si le document est considéré comme pertinent:

- On calcule (42) une fonction d'ambiguïté de ce concept (12) dans ce document, distincte de la fonction de pertinence, ce calcul étant estimé à partir de la présence dans le document de différents sens de ce concept, et

- On associe (44) un score d'ambiguïté au document considéré comme pertinent.

2. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon la revendication 1, dans lequel la fonction de pertinence mesure la présence du concept (12) et des concepts du voisinage sémantique (16) de ce concept (12) dans le document.

3. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon la revendication 1 ou 2, dans lequel, le voisinage sémantique de ce concept (12) comportant plusieurs nuages sémantiques (16) de sens distincts, la fonction d'ambiguïté compare la présence de concepts (12) appartenant à un nuage sémantique (16) correspondant à un sens prédéterminé du concept (12) avec la présence de concepts appartenant à des nuages sémantiques (16) différents.

4. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon la revendication 3, dans lequel la présence de chacun des concepts appartenant aux différents nuages sémantiques (16) est pondérée par un coefficient prédéterminé.

5. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon l'une quelconque des revendications 1 à 4, comportant une étape préalable de détection de concepts ambigus (18), c'est à dire de concepts comportant plusieurs nuages sémantiques (16) de sens différents dans leur même voisinage sémantique.

6. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon la revendication 5, dans lequel, lors de l'étape de détection préalable, deux concepts sont considérés comme ambigus (18A) s'ils sont reliés entre eux par au moins deux liens sémantiques (14) différents.

7. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon la revendication 5 ou 6, dans lequel, lors de l'étape de détection préalable, un concept est considéré comme ambigu (18B) s'il est relié à au moins deux nuages sémantiques (16) de sens différents.

8. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon l'une quelconque des revendications 5 à 7, dans lequel, le concept (12) appartenant à une base de connaissance (10) obtenue par fusion d'une première base de connaissances (10A) avec une seconde base de connaissances (10B), l'étape préalable de détection des concepts ambigus est réalisée lors de la fusion.

9. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon la revendication 8, dans lequel, lors de l'étape de détection des concepts ambigus, un concept de la première base de connaissances (10A) est considéré comme ambigu (18C) s'il est relié par un nouveau lien à un autre concept de la première base de connaissances (10A).

10. Procédé d'estimation de la pertinence d'un document par rapport à un concept (12) selon la revendication 8 ou 9, dans lequel, lors de l'étape de détection des concepts ambigus, un concept de la première base de connaissances (10A) est considéré comme ambigu (18C) s'il est relié à au moins un nuage sémantique de la seconde base de connaissances (10B).